[Published as a chapter in *Reduction*: *Between the Mind and the Brain*, Hieke, A. and H. Leitgeb (eds.), Ontos Verlag, pp. 117-133, 2009.]

THE ELIMINATION OF MEANING IN COMPUTATIONAL THEORIES OF MIND

PAUL SCHWEIZER
University of Edinburgh

Abstract
According to the traditional conception of the mind, semantical content is perhaps the most important feature distinguishing mental from non-mental systems. And this traditional conception has been incorporated into the foundations of contemporary scientific approaches to the mind, insofar as the notion of 'mental representation' is adopted as a primary theoretical device. Symbolic representations are posited as the internal structures that carry the information utilized by intelligent systems, and they also comprise the formal elements over which cognitive computations are performed. But a fundamental tension is built into the picture - to the extent that symbolic 'representations' are formal elements of computation, their alleged content is completely gratuitous. I argue that the computational paradigm is thematically inconsistent with the search for content or its supposed 'vehicles'. Instead, the concern of computational models of cognition should be with the *processing structures* that yield the right kinds of input/output profiles, and with how these structures can be implemented in the brain.

1. THE COMPUTATIONAL PARADIGM

According to the traditional conception of the mind, semantical content is perhaps the most important feature distinguishing mental from non-mental systems. For example, in the scholastic tradition revived by Brentano (1874), the *essential* feature of mental states is their 'aboutness' or intrinsic representational aspect. And this traditional conception has been incorporated into the foundations of contemporary scientific approaches to the mind, insofar as the notion of 'mental representation' is adopted as a primary theoretical device. For example, in classical (e.g. Fodorian) cognitive science, Brentano's legacy is preserved in the view that the properly cognitive level is distinguished precisely by appeal to representational content. There are many different levels of description and explanation in the natural world, from quarks all the way to quasars, and according to Fodor,

it is only when the states of a system are treated as representational that we are dealing with the genuinely cognitive level.

The classical paradigm in cognitive science derives from Turing's basic model of computation as rule governed transformations on a set of syntactical elements, and it has taken perhaps its most literal form of expression in terms of Fodor's Language of Thought hypothesis (Fodor 1975, 2008) (henceforward LOT), wherein mental processes are explicitly viewed as formal operations on a linguistically structured system of internal symbols. So in the present discussion I will use the LOT as a very clear exemplar of the classical approach, although the basic points generalize far beyond Fodor. According to the LOT, propositional attitude states, such as belief and desire, are treated as computational relations to sentences in an internal processing language, and where the LOT sentence serves to represent or encode the propositional content of the intentional state. Symbolic representations are thus posited as the internal structures that carry the information utilized by intelligent systems, and they also comprise the formal elements over which cognitive computations are performed. According to the traditional and widely accepted belief-desire framework of psychological explanation, an agent's actions are both *caused* and explained by intentional states such as belief and desire. And on the LOT model, these states are sustained via sentences in the head that are formally manipulated by the cognitive processes which lead to actions.

Fodor notes that particular tokens of these LOT sentences could well turn out to be specific neuronal configurations or brain states. The formal syntax of LOT thus plays a crucial triad of roles: it can represent meaning, it's the medium of cognitive computation, and it can be physically realized. So the syntax of LOT can in principle supply a link between the high level intentional description of a cognitive agent, and the actual neuronal process that enjoy causal power. This triad of roles allows content bearing states, such as propositional attitudes, to explain salient pieces of behavior, such as bodily motions, if the intermediary syntax is seen as realized in neurophysiological configurations of the brain. Because the tokens of LOT are semantically interpretable and physically realizable, they form a key theoretical bridge between content and causation. In this manner, a very elegant (possible) answer is supplied to the longstanding theoretical question of how mental states individuated in terms of their content, such as beliefs and desires, could be viewed as causes of actual behaviour, without violating fundamental conservation laws in physics.

So at first sight, this computational approach to cognition might seem to provide a compelling and harmonious theory of the mind/brain, potentially uniting the traditional notion of mental representation with the causally efficacious level of neural machinery. But alas, a fundamental tension is already built into the picture: a central purpose of the symbolic structures is to carry content, and yet, to the extent that they are formal elements of computation, their alleged content is completely gratuitous. Computation is essentially a series of manipulations performed on *uninterpreted* syntax, and formal structure alone is sufficient for all effective procedures. The specification and operation of such procedures makes no reference whatever to the intended meaning of the symbols involved. Indeed, it is precisely this limitation to syntactic *form* that has enabled computation to emerge as a mathematically rigorous discipline. If syntax alone is not sufficient, and additional understanding or interpretation is required, then the procedure in question is, by definition, *not* an effective one. But then the purported content of mental 'representations' is rendered superfluous to the computations that comprise the 'cognitive' processes of cognitive science. The intended interpretation of internal syntax makes absolutely no difference to the formal mechanics of mind.

2. The Connectionist Alternative

For a number of years now there has been a high profile struggle between opposing camps within the computational approach to the mind. In contrast to the classical paradigm derived from Turing, connectionist systems are based on networks of large numbers of simple but highly interconnected units that are brain-like in their inspiration. But according to Fodor (and Pylyshyn 1988), the brain-like architecture of connectionist networks tells us nothing about their suitability as models of *cognitive* processing, since it still leaves open the question of whether the mind is such a network at the *representational* level. He concedes that the connectionist approach may be the right type of architecture for the medium of implementation, which would mean that it characterizes a level below that of genuine mental structure. In view of the foregoing tension within the classical paradigm concerning formal syntax and the inefficacy of content, I would argue that Fodor is on the wrong track when he insists that, within a computational approach, the representational level is fundamental. Instead, I would argue that the internal processing structures yielding the salient input/output profiles are all that matter, whether or not these are thought of as content bear-

ing. However, a number of connectionists have taken up Fodor's challenge and seek out ways of projecting representational content onto artificial neural networks.

One comparatively recent such attempt (Churchland 1988, Laakso, A. and G. Cottrell 2000, O'Brien, G. and J. Opie 2001) uses cluster analysis to locate 'vehicles' of representational content within artificial neural networks, where such clusters serve as surrogates for the classical notion of internal syntax. Along with serious difficulties in equating clusters with the syntax of traditional computation, I would contend that such attempts suffer from exactly the same built-in tension that afflicts the LOT model; namely, the purported content for which the clusters serve as vehicles does no work in the processing path leading from inputs to outputs. Just as in the classical case, the postulation of content within the connectionist framework is gratuitous, because it plays no role in the cognitive manipulation of inputs to yield the salient outputs. Indeed, if content weren't gratuitous, then computational versions of cognitive processing would be lamentably deficient in terms of their specification of the inputs. These are characterized solely in formal or syntactical terms, and content is entirely absent from the external stimuli recognized by the operations that can be defined within the model. If representational content were at all relevant, then cognitive systems would have to process content *itself*. But according to computational methods, content is not specified with the input, nor does it play any efficacious role in internal processing. So, from a perspective that takes computation as the theoretical foundation for cognition, it seems quite retrograde to posit content on top of the factors that do the actual work. Surely this is an ideal occasion for employing Ockham's razor.

3. The Chinese Room Argument

Of course, John Searle's (1980) celebrated Chinese Room Argument (henceforward CRA) runs the dialectic in exactly the reverse direction: rather than taking the formal, syntactic nature of computation as a reason for eschewing content in a properly naturalistic approach to the mind, Searle instead takes it as a reason for rejecting computation as the appropriate theory of the mental.

So, from the perspective of the present discussion, it is instructive to explicitly cast Searle's argument in terms of the separability of syntactical structure from its intended meaning. In what follows I will abstract away

from the somewhat picturesque details of Searle's original version and express the logical core of the CRA via two premises and a conclusion:

> (1) semantical content is an essential feature of the mind,
> (2) syntactical manipulations cannot capture this content, therefore
> (3) the mind cannot be reduced to a system of syntactical manipulations.

Preimse (1) is an expression of the traditional conception of mentality, and is accepted by both Searle and by his opponents in orthodox cognitive science and AI. As stated above, classical cognitive science and AI view the mind according to the model of rule governed symbol manipulation, and premise (1) is embraced insofar as the manipulated symbols are supposed to possess representational content. Searle's dispute with cognitive science and AI centers on his rejection of the idea that internal computation can shed any real light on mental content, which leads to his conclusion (3), and to a concomitant dismissal of the research paradigm central to cognitive science and AI.

In response, a standard line for defenders of the paradigm is to try and defuse the CRA by arguing against premise (2), and claiming that the manipulated symbols really do possess some canonical meaning or privileged interpretation. However, I would urge that this is a strategic error for those who wish to defend the computational approach. As stated above, a distinguishing mathematical virtue of computational systems is precisely the fact that the formal calculus can be executed without any appeal to meaning. Not only is an interpretation intrinsically unnecessary to the operation of computational procedures, but furthermore, there is no unique interpretation determined by the computational syntax, and in general there are arbitrarily many distinct models for any given formal system.

Many classical *negative* results in mathematical logic stem from this separability between formal syntax and meaning. The various upward and downward Löwenheim-Skolem theorems show that formal systems cannot capture intended meaning with respect to infinite cardinalities. As another eminent example, Gödel's incompleteness theorems involve taking a formal system designed to be 'about' the natural numbers, and systematically reinterpreting it in terms of its own syntax and proof structure. As a consequence of this 'unintended' interpretation, Gödel is able to prove that arithmetical truth, an exemplary *semantical* notion, cannot, in principle, be captured by finitary proof-theoretic means.

Computational formalisms are syntactically closed systems, and in this regard it is fitting to view them in narrow or solipsistic terms. They are, by their very nature, independent of the 'external world' of their intended meaning and, as mentioned above, they are incapable of capturing a unique interpretation, since they cannot distinguish between any number of alternative models. This can be encapsulated in the observation that the relation between syntax and semantics is fundamentally *one-to-many*; any given formal system will have arbitrarily many different interpretations (in the very strongest case, a 'categorical' theory can determine its models up to isomorphism). And this intrinsically one-to-many character obviates the possibility of deriving or even attributing a unique semantical content merely on the basis of computational structure.

These (and a host of other) powerful results on the inherent limitations of syntactical methods would seem to cast a rather deflationary light on the project of explicating *mental content* within a computational framework. Indeed, they would seem to render such goals as providing a computational account of natural language semantics or propositional attitude states profoundly problematic. Non-standard models exist even for such rigorously defined domains as first-order arithmetic and fully axiomatized geometry. And if the precise, artificial system of first-order arithmetic cannot even impose isomorphism on its various models, how then could a *program*, designed to process a specific natural language, say Chinese, supply a basis for the claim that the units of Chinese syntax posses a *unique* meaning?

So I think that the advocates of computationalism make the wrong move by accepting Searle's bait and taking on board the seemingly intractable 'symbol grounding problem' that results. Instead I would accept Searle's negative premise (2) and agree that computation is too weak to underwrite any interesting version of (1). Hence I would concur with Searle's reasoning to the extent of accepting the salient *conditional* claim that *if* (1) is true *then* (3) is true as well. So the real crux of the issue lies in the truth-value of (1), without which the consequent of the *if-then* statement cannot be detached as a free-standing conclusion. Only by accepting the traditional, *a priori* notion of mentality assumed in premise (1), does (3) follow from the truth of (2). And it's here that I would diverge from the views of both Searle and orthodox cognitive science.

4. Conscious Presentation

In explicating and defending his pivotal premise (1), Searle (1990, 1992) again follows Brentano, in claiming that the human mind possesses original intentionality because it can experience conscious presentations of the objects that its representational states are 'about'. Thus it is conscious experience that ultimately underwrites the intrinsic aboutness of genuine intentional states. So Searle holds that consciousness supplies the basis for the truth of premise (1), and he further believes that consciousness arises from the specific causal powers of the brain considered as a physical structure, rather than via the implementation of some abstract 'formal shadow', be it classical or connectionist. Hence intentionality is tethered to brain processes via consciousness, and Searle thereby attempts to naturalize the traditional notion of mentality, while at the same time discrediting the computational paradigm, since he argues that computation has nothing to do with consciousness.

And while I would again agree with Searle's view that consciousness arises from physical brain activities rather than from multiply realizable computational structure, I would nevertheless argue, contra Searle, that conscious experience, just like symbol manipulation, is too weak to underwrite any interesting version of tenet (1). With respect to the view that conscious experience is the cornerstone of intentionality, the CRA simply begs the question, because it presupposes that the homunculus Searle, replete with conscious presentations, *really does* understand English in some special way. Searle appeals to himself as the locus of genuine intentionality in the Chinese Room, and he would support this by citing the fact that he is consciously aware of the meanings of English expressions. For example, he can entertain a conscious image of the referent of the English string 'h-a-m-b-u-r-g-e-r', while for him the strings of Chinese characters are completely devoid of conscious meanings. Ostensibly, this special understanding of English enables him to follow the program and manipulate the 'meaningless' Chinese symbols. Hence lack of conscious presentation with respect to the semantics of Chinese constitutes the real asymmetry between the two languages, and this underlies Searle's claim that genuine understanding occurs in the case of one language and not the other.

But this line of thought is not particularly compelling, since one can easily concede that Searle has episodes of conscious awareness which attend his processing of English, while at the same time denying that these episodes are sufficient to establish intrinsic content, or to ground the semantics of natural language expressions. Indeed, the mere occurrence of conscious presentations is too weak to even establish that they themselves

play a role in Searle's ability to follow the English instruction manual. Instead, I would argue that what consciousness actually provides is the foundation for the subjective *impression*, had by Searle and others, that the human mind enjoys some mysterious and seemingly magical form of intentionality with the power to uniquely determine representational content.

Thus when Searle contends that our mental states are 'really about' various external objects and states of affairs, this is merely an expression of the fact that, introspectively, it *seems to us* as if our mental states had some such special property. As argued in (Schweizer 1994), conscious experience is clearly sufficient to provide the source for this belief, since conscious experience intrinsic to how (some of) our mental states appear to us. But it cannot provide a basis for concluding that the belief is *true*, unless consciousness is something much more mysterious and powerful than the resources of natural science can allow. Brentano famously dismissed naturalism, and he thereby gave himself some room for the claim that consciousness underwrites the mind's essential intentionality. However, if one accepts naturalism and deems consciousness to be a phenomenon supported by, say, the causal properties of electrochemical reactions taking place inside the skull, then one should just bite the bullet and accept that it is too weak to support Brentano's thesis that intentionality is an essential feature of the mind.

It would be straying too far from the main goal of the article to expand on this latter claim at any great length, but considerations based on the 'narrow' status of consciousness should suffice to illustrate the central point. It is widely held by naturalistically inclined philosophers that psychological states and properties must supervene upon occurrent, *internal*, physical states and processes of organisms. This principle of 'psychological autonomy' should clearly apply to conscious states as well, and as a consequence, factors outside the boundaries of an organism cannot affect consciousness, unless they make some relevant impact on the occurrent, internal physical states and processes of that organism, most typically through inputs to the sensory mechanisms. But then the objection raised by Searle in the CRA against the computational paradigm comes back to undermine his own position: the intrinsic relation between consciousness and its object becomes one-to-many, just as the relation between computational syntax and its interpretation is one-to-many. Any number of different, non-standard, causes can yield exactly the same conscious experience (by inducing exactly the same internal physical states and processes), just as a given formal system can have arbitrarily many distinct interpretations.

Therefore conscious experience is, by its very nature, too weak to determine a unique external object that one is conscious of. This problem is at the heart of Cartesian scepticism, and it still remains firmly entrenched within the narrow confines of naturalism. According to Descartes there could be any number of different 'causal' circumstances correlated with the same conscious state, and he therefore entertained a very radical version of the one-to-many problem, in which even a malignant demon could not be ruled out as a non-standard model. In a more contemporary guise, Putnam's (1981) celebrated brains-in-a-vat argument exploits this solipsistic feature to show that conscious psychological states are too weak to capture the semantics of natural language.

5. Representation as Heuristics

There have been a number of high profile positions advanced in negative reaction to 'traditional' cognitive science that take anti-representationalism as one their hallmarks, including dynamical systems theory (e.g Van Gelder 1996), behaviour based robotics (e.g. Brooks 1996), approaches utilizing sensory-motor affordances (e.g. Noë 2004), and some varieties of connectionism (that deliberately refuse Fodor's challenge). A common factor is that these views all advance some version of the slogan 'intelligence without representation'. In order to locate my position on the salient philosophical landscape, it is worth noting that it is *not* anti-representational in this sense. On my view, there could well be internal structures that play many of the roles that people would ordinarily expect of representations, and this is especially true at the level of perception, sensory-motor control and navigation. So I would be quite happy to accept things like spatial encodings, somatic emulators, internal mirrorings of relevant aspects of the external environment. Ultimately this boils down to questions that must be settled empirically in the case of biologically induced agents, but unlike the anti-representationalists, I do not deny that the most plausible form of cognitive architecture may well incorporate internal structures and stand-ins that many people would be tempted to *call* 'representations'.

But I would argue that this label should be construed purely in a weak, operational sense, and should not be conflated with the more robust traditional conception. To the extent that internal structures can encode, mirror or model external objects and states of affairs, they do so via their own causal and/or syntactic properties. And again, to the extent that they influence behaviour or the internal processing of inputs to yield outputs, they do

this solely in virtue of their causal and/or syntactic attributes. There is nothing about these internal structures that could support Searle's or Brentano's notion of original intentionality, and there is no independent or objective fact of the matter regarding their 'real' content or meaning.

And similarly, my view is not eliminativist in the sense of Churchland (1981), because, just as in the case of low level activities such as sensation, navigation and motor control, so too in the case of higher level activities such as rational deliberation and the interaction of propositional attitudes – my position is not based on conjectures about the non-existence of various internal elements as revealed by future scientific research. Maybe it will turn out to be a theoretically fruitful level of description to view the brain as implementing a full blown system of recursive syntax. So, I would not deny, in advance of weighty empirical evidence, that there *may* even be processing structures that play the role of Fodor's belief and desire boxes, internal sentences, etc. (although I *would* find this rather surprising). So I would not at this point rule out the possibility that there may be some type of *operational* reduction of traditional psychological concepts to functional or neurophysiological states that could prove useful in predicting behaviour (see Schweizer 2001 for more discussion). Instead, my point is that even if there were such neural structures implementing an internal LOT, this still wouldn't ground traditional semantics and genuine aboutness. As will be argued in more detail in the next section, these structures would have the relevant causal/syntatctic properties but not the semantic ones.

So what I deny is not that there may be internal mechanisms that reflect external properties and states of affairs in systematic and biologically useful ways. Instead I would deny that there is anything *more* to this phenomenon than highly sensitive and evolved relations of calibration between the internal workings of an organism and its specialized environmental context. Evolutionary history can be invoked to yield interesting heuristics with respect to these mechanical relations of calibration, and perhaps support counterfactuals regarding their role in the organism's adaptive success. But evolution is based on random mutation, and natural 'selection' is an equally purposeless mechanism. Neither can provide the theoretical resources sufficient to ground the strong traditional notion of 'genuine aboutness'.

Thus if I had to coin a competing slogan to encapsulate my own position, it would be something like 'representation without intentionality'. If one is truly committed to naturalism, then there is only a difference of degree and complexity but not in kind between, say, the reflection of moon-

light in a pond and the retinal image of the moon in some organism's visual system. Proponents of the orthodox view are inclined to think that a sufficient difference in degree and complexity somehow yields an esoteric difference in *kind*, a difference that allows us to cross the conceptual boundary from mere causal correlations to 'genuine aboutness'. But I would contend that naturalism itself supplies an asymptotic limit for this curve, and that the boundary can be crossed only by invoking non-natural factors.

6. Behavior versus Meaning

The considerations presented so far have been motivated within the framework of a computational approach to the mind. But mental processes and natural language semantics clearly have many intimate philosophical connections, and the foregoing one-to-many relation underlying the symbol grounding problem has well known consequences for the linguistic theory of meaning. If one accepts the allied principle of psychological autonomy, then it follows that the mind is too weak to determine what its internal components are 'really about', and this extends to the case of expressions in natural language as well. The famed conclusion of Putnam's Twin Earth argument (Putnam 1975) is that "meanings ain't in the head", and this is because narrow psychological states are incapable of determining the reference relation for terms in our public languages. But rather than abandon natural language semantics in light of the problem, the externalist quite rightly abandons the traditional idea that the intentionality of mental states provides the foundation for linguistic reference.

Putnam's strategy is to directly invoke external circumstances in the characterization of meaning for natural languages. The externalist approach exploits direct, ostensive access to the world, thus circumventing the difficulty by relieving mental states of their referential burden. On such an approach, the object of reference can only be specified by indexical appeal to the object itself, and in principle it *cannot* be determined merely from the psychological states of the language user. Direct appeal to the actual environment and linguistic community in which the cognitive agent is situated then plays the principal role in determining the match-up between language and world. Putnam's strategy offers a viable account of linguistic reference *precisely because* it transgresses the boundaries of the mind intrinsic to the explanatory project of cognitive science. The externalist must invoke broad environmental factors, since nothing internal to a cognitive system is capable of uniquely capturing the purported 'content' of its rep-

resentations and thereby semantically grounding its internal states. And from this it follows that original content is not a property of the representation *qua* cognitive structure, and hence it is not the cognitive structure itself that provides the theoretical basis for meaning. Indeed, outside factors then do the real work, and the purported *semantical* aspect of internal configurations is trivialized.

However, in normal, everyday practice, we continually use sentences of public language to ascribe various content bearing mental states, both to ourselves and others, and it is here that a potential confusion arises. A defender of the tradition might argue that the *truth* of such ascriptions shows that there is still a legitimate fact of the matter regarding mental content, and hence that there is an objective match-up problem remaining to be solved. When an agent is correctly attributed a given propositional attitude, such as the belief that φ, this captures an actual feature of their doxastic configuration and must be supported by some corresponding aspect of their internal make up.

At this point I do not wish to become embroiled in the 'Folk Psychology' debate, but in terms of the present discussion it is important to note that such a line of argument makes an unwarranted extrapolation from our common sense practices, because the age-old customs of folk psychology are independent of any assumptions about internal symbols, states or structures. Observable behavior and context are the relevant criteria, and the truth-conditions for such ascriptions are founded on external, macroscopic and operational considerations. As in everyday life, one can use behavioral and environmental factors to adduce that, say, Jones believes that lager quenches thirst, but this practice makes no assumptions about the nature or even existence of an internal representation encoding the propositional content of the belief . The attribution concerns Jones as an unanalyzed unit, a black box whose actions take place within a particular environmental and linguistic setting. It gives no handle whatever on postulating hidden internal cogs and levers that generate Jones' actions, and it's perfectly compatible with an agnostic disregard of such inner workings.

At this stage, an ardent representationalist is likely to invoke the belief-desire framework of psychological explanation to defend a realist account of internal meaning. As mentioned at the start of the paper, not only do we ascribe various content bearing states to ourselves and others, but furthermore we habitually *use* such ascriptions to explain and successfully predict behavior. According to this widely accepted framework, psychological states individuated in terms of their *content*, such as beliefs and de-

sires, are *causally* responsible for a host of rational actions. Hence, it might be argued, the belief-desire framework can successfully predict behavior from the outside, precisely because it mirrors the internal processing structure that causes the behavior.

Thus when, from the outside, we justifiably ascribe to Jones the belief that lager quenches thirst, Fodor would have it that a token of some mentalese sentence, say 'n%^7 £#~ %&!+', which encodes the same semantical content as the English ascription, has been duly etched into her 'belief box'. This physical implementation of mentalese syntax is then poised to interact with other physically implemented tokens in her desire box to produce assorted forms of rational action, such as standing up and reaching for a pint. In this manner, the truth of propositional attitude ascriptions is directly correlated with salient internal configurations of the agent.

But this purported correlation breaks down at its most vital point – the level of semantical content. For the story to work, the sentences 'lager quenches thirst' and 'n%^7 £#~ %&!+' must both express the same proposition. Yet as a medium of classical computation, the LOT is just a scheme for rule governed symbol manipulation. Syntax churning within a formal system is fundamentally different from the operation of a public language, and it is a significant conflation to impute to the former the same semantical properties conventionally attributed to the latter. English is acquired and exercised in an inter-subjectively accessible context with which the entire sociolinguistic community has indexical contact. There are shared criteria for the correct use of natural language sentences and the rules under which various expressions are deployed, and there are direct, ostensive ties between publicly produced syntactic tokens and their referents. In vivid contrast, there are no such shared criteria nor public ties for the hidden, internal sentences of mentalese. The LOT serves as an extreme example of a *private* language (Wittgenstein, 1953), and as such it has no communal truth conditions nor standard semantic properties. Indeed, the LOT is so private it's even hidden from the introspective awareness of the individual agent, and it thereby also eludes Searle's traditional association of linguistic meaning with agent-based intentionality.

As elements in a formal system, there is no fact of the matter concerning what the internal sentences of mentalese 'really mean'. At best, these conjectured tokens of computational syntax would successfully govern our behavior in familiar surroundings, but they would fail to do so if we were placed in radically different circumstances. So they are merely

calibrated with the environment in which they happened to develop, and this historical fact is not sufficient to imbue them with objective content. To the extent that these hypothetical symbols successfully govern behavior, they do so purely in terms of their formal, syntactical properties, and as noted before, there is no work left to be done by their intended interpretation. On a computational approach to the mind, it is processing structure and not semantics that is the cause of human action.

So at this point a wedge must be driven between two apparently related but nonetheless quite distinct theoretical projects. There is very significant difference between a theory of natural language semantics and a psychological theory regarding the internal states causally responsible for our input/output profiles. The former is an idealized and normative endeavor, concerned with articulating high level characterizations which reflect the socially agreed truth-conditions for sentences in a public language. As such, this endeavour has no direct bearing on an essentially descriptive account of the internal mechanisms responsible for processing cognitive inputs and yielding various behavioural outputs, even when we consider the production of *verba*l behaviour, or the common sense attribution of various propositional attitude states *using* natural language.

Hence I would diagnose the classical Fodorian effort to build semantical content into a computational theory of mind as an infelicitous failure to separate these two projects at exactly the point where they should *not* coalesce. The infelicity of this move is already apparent in *Psychosemantics*, where Fodor (1987) tries to address the notorious problems of wide versus narrow content introduced by Putnam and later Burge (1979). In an attempt to defend his narrow version of content against Twin Earth objections, Fodor is forced to claim that "… what my water-thoughts share with Twin 'water'-thoughts *isn't* content. Narrow content is radically inexpressible, because it's only content *potentially*;" (p. 50). But this sounds uncomfortably close to equivocation, and invites the question – why call it 'content' *at all*? Fodor goes on to say that "a narrow content is *essentially* a function from contexts onto truth conditions;" (p. 53), so that in the context of Earth this function yields thoughts about H_2O, and on Twin Earth it yields thoughts about XYZ. He states that this abstract function is implemented in the human brain, whereby it enjoys causal efficacy in the physical world. But it's crucial to note that the distinguishing characteristics of such abstruse functions are woefully underspecified by the brute facts of physical brain structure and natural selection. Mere terrestrial teleology is

one thing, but how on earth could biological evolution select a function designed to yield XYZ thoughts on another planet?

This account appears to be a strained attempt to appropriate and internalize a normative, idealized position in the theory of natural language semantics, rather than to provide a naturalistically plausible story about cognitive processing. Instead of narrow 'content', what such Twins have in common is the same internal processing structure, and this produces the same outputs when given the same inputs, regardless of the input's distal source in the environment. In contrast to Fodor's claim quoted above concerning the essential nature of narrow content, the proper domain of the implemented cognitive function is *inputs* and *not contexts*, and this is precisely why individual cognitive systems cannot capture the semantics of public languages.

7. Conclusion

According to the position advocated herein, the traditional commitment to representational *content* constitutes a retrograde step within the context of naturalistic explanation. The crucial point to notice is that internal 'representations' do all their scientifically tangible *cognitive* work solely in virtue of their physical/formal/mathematical structure. There is nothing about them, qua efficacious elements of internal processing, that is 'about' anything else. Content is not an explicit component of the input, nor is it acted upon or transformed via cognitive computations. All that is explicitly present and causally relevant are computational structure plus supporting physical mechanisms, which is exactly what one would expect from a naturalistic account.

In order for cognitive structures to do their job, there is no need to posit some additional 'content', 'semantical value', or 'external referent'. Such representation talk may serve a useful heuristic role, but it remains a conventional, observer-relative ascription, and accordingly there's no independent fact of the matter, and so there isn't a sense in which it's possible to go wrong or be mistaken about what an internal configuration is 'really' about. Instead, representational content can be projected onto an internal structure when this type of gloss plays an opportune role in characterizing the overall processing activities which govern the system's interactions with its environment, and hence in predicting its salient input/output patterns. But it is simply a matter of convenience, convention and choice, and

does not reveal an underlying fact of the matter nor any essential characteristics of the system.

From the point of view of the system, these internal structures are manipulated *directly*, and the notion that they are 'directed towards' something else plays no role in the pathways leading from cognitive inputs to intelligent outputs. Hence the symbol grounding problem is a red herring – it isn't necessary to quest after some elusive and mysterious layer of 'real' content, for which these internal structures serve as the mere syntactic vehicle. Syntactical and physical processes are all we have, and their efficacy is not affected by the purported presence or absence of meaning. I would argue that the computational paradigm is thematically inconsistent with the search for content or its supposed 'vehicles'. Instead, the concern of computational models of cognition should be with the internal *processing structures* that yield the right kinds of input/output profiles of a system embedded in a particular environmental context, and with how such processing structures are implemented in the system's physical machinery. These are the factors that do the work and are sufficient to explain all of the empirical data, and they do this using the normal theoretical resources of natural science. Indeed, the postulation of content as the essential feature distinguishing mental from non-mental systems should be seen as the last remaining vestige of Cartesian dualism, and, contra Fodor, naturalized cognition has no place for a semantical 'ghost in the machine'. When it comes to computation and content, only the vehicle is required, not the excess baggage.

References

Brooks, R. 1996 "Intelligence without Representation" in *Mind Design II*, J. Haugeland (ed.), MIT Press.

Brentano, F. 1874 *Psychology from an Empirical Standpoint*.

Burge, T. 1979 "Individualism and the Mental", in French, P., Euhling, T., And Wettstein, H. (eds.), *Studies in Epistemology*, vol. 4, *Midwest Studies in Philosophy*, University of Minnesota Press.

Churchland, P.M. 1981 "Eliminative Materialism and the Propositional Attitudes", The Journal of Philosophy 78: 67-90.

Churchland, P.M. 1998 "Conceptual Similarity Across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered", *Journal of Philosophy*, 96(1): 5-32.

Fodor, J. 1975 *The Language of Thought.* Harvester Press.

Fodor, J. 1987 *Psychosemantics*, MIT Press.

Fodor, J. 2008 LOT 2 *The Language of Thought Revisited*, Oxford University Press.

Fodor, J. and Z. Pylyshyn 1988 “Connectionism and Cognitive Architecture: A Critical Analysis”, *Cognition,* 28: 3-71.

Laakso, A. and G. Cottrell 2000 “Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems”, *Philosophical Psychology*, 13(1): 47-76.

Noë, A. 2004 *Action in Perception*, MIT Press.

O’Brien, G. and J. Opie 2001 “Connectionist Vehicles, Structural Resemblance, and the Phenomenal Mind”, *Communication and Cognition*, 34: 13-38.

Putnam, H. 1975 “The Meaning of ‘Meaning’”, in *Mind, Language and Reality*, Cambridge University Press.

Putnam, H. 1981 “Brains in a Vat”, in *Reason, Truth and History*, Cambridge University Press.

Schweizer, P. 1994 “Intentionality, Qualia and Mind/Brain Identity”, *Minds and Machines* 4: 259-282.

Schweizer, P. 2001 “Realization, Reduction and Psychological Autonomy” *Synthese* 126: 383-405.

Searle, J. 1980 “Minds, Brains and Programs”, *Behavioral and Brain Sciences*, 3: 417-424.

Searle, J. 1990 “Consciousness, Explanatory Inversion and Cognitive Science”, *Behavioral and Brain Sciences*, 13: 585-596.

Searle, J. 1992 *The Rediscovery of the Mind*, MIT Press.

Van Gelder, T. 1996 “Dynamics and Cognition” in *Mind Design II*, J. Haugeland (ed.), MIT Press.

Wittgentein, L. 1953 *Philosophical Investigations*.